METHODS OF ASSESSING THE ACHIEVEMENT OF STUDENTS IN CHARTER SCHOOLS

Caroline M. Hoxby
Harvard University

Sonali Murarka
National Bureau of Economic Research

Prepared for the National Conference on Charter School Research
at Vanderbilt University on September 28, 2006

NATIONAL CENTER ON
School Choice

This working paper is supported by the National Center on School Choice, which is funded in part by the Department of Education's Institute of Education Sciences (R305A040043). For more information, please visit the National Center on School Choice's website at http://www.vanderbilt.edu/schoolchoice/.

METHODS OF ASSESSING THE ACHIEVEMENT OF STUDENTS IN CHARTER SCHOOLS[1]

Caroline Hoxby[2] and Sonali Murarka[3]

I. A Challenge for Assessment

Policy makers wish to have evidence on the effect that charter schools have on their students' achievement. Such evidence has two different potential uses. First, a policy maker who is considering expanding or contracting the availability of charter schools in his state may wish to know whether parents' desire to send their children to charter schools is based on their observations of achievement or based on other criteria. That is, a policy maker may say to himself, "Unless it can be shown that charter schools have X effect on achievement, I do not want to allow students to attend them even if the students and their parents desire it." (The

---

[1] Paper prepared for the conference entitled "Charter Schools? What Fosters Growth and Outcomes?", sponsored by the National Center on School Choice at Vanderbilt University on 28-29 September 2006.. The authors wish to thank the Chicago Consortium on School Research, the Chicago Public Schools, the Chicago International Charter School, the New York City Center for Charter School Excellence, and the New York Department of Education for data, information, cooperation, and a great deal of help. None of these organizations or their staff members are, however, responsible for the content of this paper. For comments that made the project stronger, the authors also wish to thank anonymous reviewers for the Institute for Education Sciences, Mark Schneider and Phoebe Cottingham of the Institute for Education Sciences (NCES and NCER, respectively), Jonah Rockoff, and seminar participants at Mathematica, Rand, UCLA, Harvard University, and the National Bureau of Economic Research.

[2] Department of Economics, Harvard University and National Bureau of Economic Research, Cambridge MA 02138. Hoxby is the corresponding author.

[3] National Bureau of Economic Research, Cambridge MA 02138.

policy maker has no question in front of him unless he is willing to impose his judgement on parents and unless there are more parents who want to send their children to charter schools than there is space available. Changing the legal availability of charter schools is a meaningless act unless there are parents who want to enroll their children in the places that become available or unavailable.) A policy maker's second potential use for evidence on charter schools' effect on achievement is very different: he may wish to discover what works in public education and he may view charter schools as laboratories in which interesting educational experiments often occur.

If the policy maker's need for evidence is of the first type, he should be supplied with evidence on how the average charter school in his state affects the achievement of students who wish to attend them. Even better, if he is considering expansion, he should be supplied with evidence on the achievement effects generated by the most-in-demand charter schools on the students who want to attend them. If he is considering contraction, he needs evidence on the achievement effects generated by the charter schools his policy is most likely to shut down.

If the policy maker's need for evidence is of the second type, he should be supplied with evidence on which charter school characteristics, if any, predict that the school has significant positive or negative effects on student achievement. This is a much more demanding evidentiary requirement because characteristics must vary substantially among charter schools, must be measurable with objective metrics, and must vary fairly independently. Intuitively, this last requirement means that clumps of characteristics are problematic. If a long school day is always found in conjunction with a long school year, guided instruction, and certain curricula, researchers will be unable to identify the independent effect of a long school day.

Because researchers find it very challenging just to produce credible estimates of the

average effect of charter schools on achievement, they have thusfar mainly attempted to generate the first type of evidence and left the second type for future studies. In this paper, we will largely follow this approach and will focus our discussion on methods for estimating the average achievement effect. We will, however, turn to the second type of evidence, briefly. in the final section.

We should say at the outset that this is primarily a paper about methods, not a paper about results. There is a need for serious discussion of methods at this time, owing the large number of studies, recently completed or in process, that purport to estimate the achievement effect of charter schools. The methods of these studies range from wholly out-of-touch with modern scientific evaluation to simple but somewhat credible to apparently sophisticated but actually misguided to rigorous and credible. It is not useful to target this paper at researchers who could be described as wholly out-of-touch with modern scientific evaluation--not only would they probably not read it, but they would also not recognize certain well-known, but modern evaluation methods. Therefore, we target this paper towards the community of serious researchers who are sincerely interested in credibly evaluating charter schools.

A final caveat is that, in no case are we interested in how charter schools might affect the achievement of students who do not want to attend them. This is because the charter school idea is inherently and deeply voluntary: there is no version of charter schools that students are forced to attend.

II. Self-Selection into Charter Schools

To the layman, it is perhaps surprising that researchers find it so difficult to answer the narrow question, "What is the effect of charter schools [in area Y, of type Z] on the achievement

of students to wish to attend them?" There are two serious problems, however: (i) selection and (ii) general equilibrium. The general equilibrium problem occurs when charter school enrollment is so pervasive that it substantially affects the peers, locations, or finances of regular public schools or private schools. That is, charter schools might affect achievement not only by their direct effect on learning but also by indirectly altering students' main alternative schooling options. In the long run, researchers will find that the general equilibrium problem is extremely hard to address. For now, charter schools account for so little enrollment that researchers can ignore the general equilibrium problem unless they are gathering evidence on one of the very few areas, such as certain counties in Arizona, where charter schools are prevalent. Unfortunately, the opposite is true of the selection problem: it not only plagues researchers now, but there is every reason to believe that is as serious in today's data as it will ever be.

In nutshell, the selection problem occurs because families self-select into charter schools for numerous reasons and many of the reasons have their own effects on achievement. Self-selection is highly prevalent at present because charter schools are an unusual, new option. Therefore, when parents submit a charter school application, they are diverging *more* from their inertial or default behavior than they will be years from now, when charter schools will less novel and probably more ubiquitous (judging from their growth thus far).

Why might students self-select into charter schools and how might the selection destroy the credibility of some evidence? A parent who switches her child to a charter school may do so because she sees that child starting to struggle academically in his regularly assigned public school. Alternatively, she might feel that the child is "getting in with the wrong crowd" or otherwise beginning to display problematic behaviors. A parent might make the switch because she wishes to give her child a clean start after a troubling event such as bullying or being taught

by a prejudiced teacher who took an unjustifiably dim view of the child's ability. Another reason for choosing a charter school may be a child's failure to make the cut for a selective magnet or exam school in the regular public system. As reasons for switching, all of the above stories, and many others like them, would lead us to expect *negative-in-trajectory* self-selection into charter schools.

To be precise, *negative-in-trajectory* self-selection occurs when a child is enrolled in a charter school because of a phenomenon that could negatively affect the child's trajectory of achievement from the time of observation onwards. For instance, having a prejudiced third grade teacher who takes an unjustifiably dim view of a child's ability because of his race might not only be expected to reduce the child's achievement in grade three but also be expected to have lingering effects. Not only is third grade knowledge useful for learning fourth grade material, but the child might internalize the teacher's prejudices and thereafter set low expectations for himself.

Self-selection into charter schools may also take a *negative-in-level* form. Negative-in-level self-selection occurs when a child is enrolled in a charter school because he has a characteristic that is predetermined at the time when we first observe him and when that characteristic reduces his level of achievement at that time but has no continuing, dynamic effects on the trajectory of his achievement. Suppose, for instance, a child has a vision problem that goes unrecognized in kindergarten and first grade. He thereafter gets corrective lenses that completely fix the problem. As a second grader, his incoming level of achievement is reduced by his (past) vision problem, but he can be expected to make regular progress in second, third, and higher grades. If this story seems contrived or if it is difficult to imagine phenomena that have one-time but no lingering, dynamic effects, this is because *in-level* selection imposes

stringent restrictions on the way in which phenomena affect achievement. Many commentators feel that these restrictions are unrealistic.

Self-selection into charter schools may also take *positive-in-trajectory* or *positive-in-level* forms. For instance, families who apply to charter schools may be those who are better able to process information about the educational system. Such information processing could be helpful in any number of ways that would improve a child's trajectory --more productive parent-teacher conferences, more accurate parent supervision of homework, and so on. Positive selection may also be triggered by an event. For example, a child may have a second grade teacher (in the regular public schools) who rates him highly or tells his family that he is under-challenged. His family may therefore seek out a charter school that promises more homework, more frequent testing, or more opportunities for advanced learning. The child's second grade experience and the family's evolving expectations for him would presumably improve his achievement regardless of his school.

Finally, there are forms of in-trajectory and in-level self-selection that are difficult to classify as positive or negative. Consider parents who disagree with the special education diagnosis or services to which their child has recently been assigned in his regular public school. They might enroll their child in a charter school in an attempt to get a fresh diagnosis. Some commentators might see this as negative self-selection: poorly informed parents trying to ignore expert opinion. Other commentators might see this as positive self-selection because only parents who are very invested in their child's education are likely to go to the trouble of contradicting the diagnosis of an expert. In any case, it is important to realize that there are plausible forms of self-selection that could lead either to upward or to downward biases in estimates of charter schools' effects. As a result, researchers must take care to eliminate all

forms of self-selection that they possibly can. This is *not* a case where, if a researcher eliminates some forms of selection bias and finds that the estimates go in a certain direction, he can reasonably project that the estimates would continue to go in that direction if he were to eliminate all forms of selection bias.

Thus far, we have provided examples of self-selection based on phenomena that are hard for researchers to observe--a bullying episode, a teacher's prejudices (either positive or negative), a family's ability to process information, a child's getting in with the wrong crowd. This is because, to the extent that selection is based on readily observable variables, researchers can statistically account for the selection so that it poses a more limited challenge to generating evidence. This point can best be illustrated with an example. For numerous reasons (cost, safety, neighborhood friends), most parents prefer not to send their young children to schools located at a great distance from their homes. It is no surprise, therefore, that the distances from a child's home to the nearest charter schools affect selection: parents tend not to apply to charter schools if the nearest ones are far away. If researchers could not observe home and school locations, this selection phenomenon would be problematic. A researcher might have a hard time explaining why children from the suburbs were not reasonable controls for central city students enrolled in a charter school. With locations being fairly observable, the problem is mitigated. The researcher may be to show how distance from a charter school affects the probability of applying. To the extent that he can model the relationship correctly or simply match students with "control" students who live at the same distance, he surmounts the challenge to evidence that is created by self-selection on the basis of distance. In short, although self-selection on the basis of observable variables occurs and requires a researcher's attention, the most troubling forms of self-selection are those that are not readily observable to the researcher.

Given that much self-selection is not readily observable and can be either negative or positive, what more can we say about it? There are three important things.

First, the more novel the charter school is, the more parents are going out of their way when they choose it. Parents who send their child to a novel school are necessarily accepting a degree of uncertainty that does not exist in their regular public school. Put another way, when a charter school starts up and has no track record, a family needs to feel more strongly about the match between its child and the school before the family is pushed over the threshold and applies. The match could be strong for a negative reason (the child needs to escape a very bad crowd) or a positive reason (the family is extremely interested in a heavier homework load). Novelty intensifies all forms of self-selection. When we say that the selection problem is intensified, we mean formally that there is a greater difference in characteristics, observed and unobserved, between those who apply to the charter school and those who do not.

Second, the later self-selection occurs in a student's career, the more information the family uses when it makes its decision. For example, a family with a child who is about to enter kindergarten knows a lot less about their child and what his experience will be in his regular public school than does a family with a child who has already attended kindergarten through fourth grades. A family cannot learn that their child is struggling academically or getting in with the wrong crowd until he does it; a family cannot have an experience with a prejudiced teacher until their child enters her classroom; a family cannot decide that the homework assigned falls short of their expectations until they see what homework is assigned. This point is obvious, but its important implication is often overlooked: the self-selection problem intensifies over a student's career simply because the family has more private information (information observable to them but not to the researcher) on which to base their decision to apply to charter school.

That is, the researcher is always at an informational disadvantage relative to the family, but the disadvantage is smallest when the child is entering kindergarten and grows dramatically over time as the family accumulates knowledge about their individual child's school experience but the researcher remains stuck with a few demographic variables, a few socio-economic variables, and perhaps some test scores (though, in a typical school, only certain grades are tested). In short, the forms of self-selection that are most worrisome intensify over time because of the growth in the information gap between researchers (who do not observe the basis for selection) and parents (who do).

Third, self-selection tends to intensify over a student's career because of the *relative* novelty of the charter school or--put another way--the relative inertial pull of the regular public school. Consider a parent with a child entering kindergarten. The charter school may be more novel than the regular public school, but their child will find either school's kindergarten quite novel. He will have to meet new peers, learn his way around a new building, adjust to a new culture, and otherwise become acclimatized. The difference in uncertainty is there, but it is less extreme than it is for, say, a prospective fourth grader. He will likely face little or no adjustment if he keeps attending his regular public school; he will face considerable novelty and acclimatization if he switches to a charter school. No reader will be surprised to hear about the relative inertial pull of the regular public school increasing over time (and being particularly high when grade-to-grade transitions occur within the same building or with the same set of peers). We are all aware of parents who would like to move for their own reasons but are wary of school switches, "save" a school switch for a year in which the child must change buildings anyway, or delay a career change until it can occur without a school switch. Yet, readers may be surprised by the serious implications of increasing relative inertia: the self-selection problem

become more and more intense over the course of a student's career and it is particularly intense between any two grades where a school switch does not occur in the regular public system.

III. Methods for Assessing Charter Schools' Effects on Achievement

In this section, we consider the three main forms of analysis that are used to evaluate the achievement of students in charter schools: comparison with controls based on observable variables, value-added analysis, and lottery-based comparisons. For each method, we ask: Which forms of selection bias does this type of analysis eliminate?

A. Comparison with Controls Based on Observable Variables

Comparison with controls based on observable variables comprises a large set of methods that can be very good at eliminating forms of selection based on observable variables. The methods under this heading (hereafter, "comparison-with-controls") include everything from fully non-parametric matching to linear multiple regression.

It is now fairly widely agreed that, within the comparison-with-controls set of methods, the most credible evidence is obtained using fairly non-parametric methods if they are feasible. These methods are most feasible if the observable variables are categorical[4] and, more importantly, if great number of student observations are available. We will start with non-parametric methods because they also have a purely expositional advantage: they clearly illustrate the key strengths and weaknesses of comparison-with-controls.

In a fully non-parametric comparison, we take each student enrolled in charter school and match him with a student who is identical to him based on all of the predetermined, observable

---

[4] Gender, for instance, is a categorical variable while family income is often a continuous variable

variables available to us. In a realistic applications, students might be put into "cells" based on their gender, race/ethnicity, free/reduced lunch status, age, grade, census block of home residence, limited English proficiency at school entry, permanent disability if any, and predetermined achievement indicators if any (*not* achievement variables that could have been affected by attending/not attending the charter school). Thus, a cell might be first graders who are black, non-Hispanic, male, free-lunch eligible, native English speakers, age six, not assigned to remedial services when evaluated at kindergarten entry, and living in Census block X. Within each cell that contains a charter school enrollee ("treatment student"), we hope to see one or more "control students" who do not attend charter school. We compute the treatment minus control difference in achievement within each cell and average across the cell differences to obtain evidence on the effect of charter schools.[5] That is, the estimated effect of charter schools is some average of within-cell differences in achievement between treatment and control students.

All other comparison-with-controls methods are essentially variants of the simple procedure described above. Without going into technicalities that are not appropriate for this paper, the other methods in this group add parametric restrictions to the procedure in the hope of gaining efficiency without introducing bias. If there are many observable variables or if some observable variables are quite continuous (family income might be measured in dollars, for instance), some cells that contain treatment students will probably not contain control students.

---

[5] See Guido W. Imbens and Jeffrey M. Wooldridge, "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature* (forthcoming) for comments on the optimality of various weighting schemes used in computing the average. Their article also provides a helpful review of several other techniques discussed here, including matching on the propensity score and local linear regression.

The researcher can overcome this problem and potentially gain efficiency by modeling the relationship between the observable variables and selection into the charter school. Such modeling may lead to nearest-neighbor and propensity score-based matching methods (where each treatment student is essentially matched with the control students who is the closest match to him based on their modeled propensity to select into charter school), a variety of regression methods (where researchers control linearly for the propensity to select into charter schools or control for a variable that is a transformation of the propensity), and methods such as local linear regression that combine matching and regression.

The advantage of comparison-with-control methods is that they can eliminate large amounts of selection bias based on observable variables--such as selection based on a child's proximity to the nearest charter schools. In addition, comparison-with-control methods can often be applied to less-than-ideal data such as achievement data available only by group--for instance, the ethnicity-by-grade-by-school-by-year groups for which reports are mandated under the No Child Left Behind act. In order to use grouped data, a researcher must find the analog to the procedure he would obtain if he performed comparison-with-controls and subsequently aggregated. This is best illustrated with an example. A charter school's grouped data might be put into cells with data for the same group from each of the regular public schools from which the charter school draws students. After achievement differences were computed based on the grouped data, a weighted average would be computed across the cells with weights based on the share of the charter school's students who were from that particular cell.

Unfortunately, there are substantial disadvantages of comparison-with-controls methods. First, they not only do not eliminate selection problems based on unobserved variables (such as a child's getting in with the wrong crowd or a family being highly motivated), they can actually

exacerbate biases associated with unobservable variables. To see this, consider two children who appear to be a perfect match for one another: they live next door; they have the same age, grade, race/ethnicity, gender, free/reduced lunch eligibility, English proficiency; and so on. One applies to charter school; the other does not. We ought to ask ourselves why two neighboring families made different decisions. It seems unlikely that the families would not communicate or would together plan to flip a coin about which child should try the charter school. Therefore, we must suppose that the two families had some reason for making different choices when they appear to have identical circumstances and are offered identical opportunities. Perhaps one child has been bullied and the neighboring child *is* the bully! More formally, by focusing on families who appear identical on observable variables but who make different schooling choices, we automatically focus on a sample of families who are disproportionately *un*alike on the unobservable variables.[6] As a result, the potential for bias from unobservable variables can grow as the researcher tightens the matching on observable variables. Indeed, in the case of the neighbors described above, a researcher might be well-advised to loosen up the match on location so that his treatment and control families are less likely to know one another. He might decide to match not on street address, but only on Census block or school attendance zone.

This leads us to the second disadvantage of comparison-with-controls methods: if they are to generate good evidence, they must be practiced by researchers who exercise considerable, good judgement. A researcher must have a deep understanding of the strengths and weaknesses of various statistical methods, and he must also have a strong intuitive sense for likely behaviors and sources of bias. In the hands of a researcher with the best training and judgement, the

---

[6] More precisely, they are disproportionately unalike on unobservable variables among families with their observable characteristics

comparison-with-controls methods tend to converge because they are deeply related. In other words, a researcher should be able to start at either end of the parametric spectrum (fully non-parametric matching or multiple linear regression) and, through specification testing and logic, arrive at similar estimation procedures. However, comparison-with-controls methods are not only not "dummy-proof," they are not even proof against modest deficiencies in training, analytic ability, data sense, or attentiveness. All too often, comparison-with-controls methods are exercised by researchers who have only the foggiest idea of the strengths and weaknesses of the method they are using and who make empirical choices out of habit or because they have seen someone else do it.

Fortunately, there is a sometimes practical way of selecting the appropriate comparison-with-controls procedure: researchers can select the one that is best at generating "gold standard" evidence when such evidence is available.[7] We shall return to this point because it is very important: the disadvantages of methods based on comparison-with-controls may be mitigated if gold standard evidence is available for a good share of the data.

B. Comparison-of-Gains-with-Controls based on Observable Variables

It is worthwhile saying something about the method of comparing achievement *gains* with controls based on observable variables. This method is often confused with value-added analysis, discussed below, in part because conventions of nomenclature are not strict. However, comparison-of-gains-with-controls is logically separate from value-added: the two methods start from fundamentally different assumptions about identification and should not be lumped

---

[7] This is the essence of the line of applied econometric research which began with the influential paper by Rejeev Dehejia and Sadek Wahba, "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94.448 (December 1999), 1053-1062.

together.

Comparison-of-gains-with-controls methods are similar to comparison-with-controls methods except that the outcome variable is a student's *gain* in achievement in a grade, rather than his level of achievement in a grade. In their most extreme form, comparison-of-gains methods are based on the assumption that a student's background affects only his initial level of achievement at school entry and that all students will thereafter make the same progress each year if given the same school treatment. Under this assumption, we should be able to compare achievement gains across all kinds of schools without worrying about whether the children have backgrounds that are at all comparable. The assumption needed for the extreme version of the method is violated by the data, however: background variables *are* correlated with rates of gain, not only with initial achievement. Indeed, the violation should be obvious to readers because some well-known achievement gaps, such as the black-white achievement gap, grow with age.[8] An achievement gap could not change systemically with age if background only affected initial achievement.

Because the extreme assumption does not hold, researchers generally use observable variables to find control students who attend regular public schools but who should, otherwise, make the same gains as students who attend charter schools. Thus, we return to the issues discussed under comparison-with-controls. Should we put students into cells and use fully non-parametric methods? Should we impose some parametric structure on the decision to select into charter schools and match on the propensity to apply? Should we use linear regression and control for the propensity to apply? The sole difference between this and the previous discussion

---

[8] See E. Meredith Phillips and Christopher Jencks, *The Black-White Test Score Gap*, Washington, DC: The Brookings Institution, 1998.

is that the outcome variable is the achievement gain.

Does using the gain reduce bias due to selection associated with unobservable variables? Does it reduce the amount of judgement required of the researcher? Arguably, even probably, yes. However, there is no *logical* reason why the problems of comparison-with-controls are reduced when we use comparison-of-gains-with-controls. For instance, there is no logical reason why a child's getting in with the wrong crowd should not affect his future rate of gain. There is no logical reason why a child from a more motivated family should not gain achievement faster, not merely start with a higher initial level of achievement.

Hereafter, we will lump comparison-of-gains-with-controls with the methods of comparison-with-controls. All of these methods have the same essential advantages and disadvantages.

C. Value-Added Analysis

Value-added analysis starts with a simple logic: if a student switches from the regular public schools to a charter school (or *vice versa*), then his own prior self can be a good control for his later self. That is, if we compare a child's achievement before and after he attends a charter school, we can potentially estimate the charter school's effect. Value-added methods obviate the necessity of our finding an appropriate control student for each treatment student because each student is his own control. This logic (which is entirely separate from comparison-of-gains, a method that does not require students to be switchers) has great intuitive appeal. Also, value-added analysis is extremely useful for the evaluation of certain educational interventions. We shall show, however, that the appeal of value-added analysis is entirely superficial when it comes to estimating charter school effects. Value-added analysis generates serious and intractable biases when used in the charter school setting.

Before delving into the nitty-gritty of value-added analysis, consider their essential logic. The fundamental assumption of value-added methods is that we can use prior information on a student to *forecast* his future achievement. Thus, we can compare a charter school student's actual achievement to the achievement we would have forecast for him had he remained in the regular public schools. We make this forecast by examining his achievement while in the regular public schools and projecting it into the future using some assumption about the relationship between past and future achievement (in the absence of school switches).

This leads us to the one major decision involved in value-added analysis: do we have sufficient information for the forecast if we use just his prior *level* of achievement, just his prior *rate of gain* in achievement, both (the prior level and rate of gain), or both interacted with some student characteristics such as race/ethnicity and free-lunch eligibility?[9] For instance, if we need only his prior rate of gain to forecast his future rate of gain, then we end up with a value-added analysis of the form:

$$(A_{it} - A_{i,t-1}) = \alpha_0 + \alpha_1 (A_{i,t-2} - A_{i,t-3}) + \alpha_2 I_{it}^{charter} + \epsilon_{it} \quad (1)$$

where A_{it} is the achievement of student i in year t, $A_{i,t-1}$ is his achievement in the previous year and so on, and $I_{it}^{charter}$ is an indicator variable for being his enrolled in a charter school in year t. We must have data on switchers to estimate this equation--otherwise, the charter school effect (α_2) is not identified. Researchers familiar with value-added analysis will observe that the above equation nests a variety of more restricted equations also based on pre-charter-school and

[9] This question should be interpreted in a strict statistical sense: which combination of variables is a sufficient statistic for the child's future achievement?

post-charter-school rates of gain. These are discussed in the footnote.[10]

Similarly, if we need (independent information) on a student's prior rate of gain and his initial level to forecast his future rate of gain, we end up with an analysis of the form:

(2) $$(A_{it}-A_{i,t-1})=\beta_0+\beta_1(A_{i,t-2}-A_{i,t-3})+\beta_2 A_{i,t_0}+\beta_3 I_{it}^{charter}+\upsilon_{it}.$$[11]

where t_0 must be prior to $t-3$. Later, we return to the question of which value-added equation to use, but we focus for now on equation (1) because this form (or one of its restricted variants) is preferred by most analysts of charter schools.[12]

How does value-added analysis help with the self-selection into charter schools that we described above? Recall that a student might switch to a charter school because he had experienced a prejudiced teacher or was getting in with the wrong crowd at his regular public

---

[10] The restricted-coefficient versions of this estimating equation are:

$$[(A_{it}-A_{i,t-1})-(A_{i,t-2}-A_{i,t-3})]=\beta_0+\beta_1(I_{it}^{charter}-I_{i,t-2}^{charter})+\varepsilon_{it}$$

and

$$[(A_{it}-A_{i,t-1})-(A_{i,t-1}-A_{i,t-2})]=\gamma_0+\gamma_1(I_{it}^{charter}-I_{i,t-2}^{charter})+\upsilon_{it}$$

These restricted specifications are nearly always rejected by basic specification tests that reject the hypothesis that the coefficient on the prior rate of gain should be one. The second of these specifications is even more likely to be rejected than the first because it suffers badly from mean-reversion: the observation for the year *t-1* is used twice, generating automatic negative serial correlation between the two gains in the presence of mean reverting test scores.

[11] When we say *both* the level and rate of gain, we mean a specification in which these variables are allowed to have effects that are independent and in which the coefficient on neither variable is restricted to being one.

[12] See, for example: Robert Bifulco and Helen F. Ladd, "The Impacts of Charter Schools on Student Achievement: Evidence from North Carolina," unpublished manuscript, April, 2004; Eric A. Hanushek, John F. Kain and Steven G. Rivkin, "The Impact of Charter Schools on Academic Achievement," Working Paper, The Cecil and Ida Green Center for the Study of Science and Society, December, 2002; and Tim Sass, "Charter Schools and Academic Achievement in Florida," unpublished manuscript, February, 2004.

school. He might switch because his motivated parents were alarmed by the light load of homework assigned in the regular school. And so on. Unfortunately, it is far from obvious that value-added analysis cures any forms of self-selection associated with unobserved variables. Suppose two children make identical gains from year *t-3* to year *t-2* but one child is pulled out of his regular public school and switched to a charter school. It is likely that there is something unobservable that is different about the two children. Yet, value-added analysis asserts that they have identical future achievement simply because they made the same gain in the most recent school year. What is to say that the switcher was not the child whose motivated parents were underwhelmed by the same gain that satisfied the other child's parents?

If value-added analysis dealt with selection bias as well as, but no better than, comparison-with-controls, we might find it appealing for its simplicity. That is, we might be grateful for the fact that the method effectively chooses the control students for us, thus eliminating an otherwise complex decision. Unfortunately, value-added analysis has problems that comparison-with-controls does not have, and--unlike the problems with comparison-with-controls, the problems with value-added analysis are intractable when it is used to evaluate charter schools.

In order to estimate equation (1), we must have data on students who switch from regular public schools to charter schools after having been tested in the regular public schools at least twice.[13] The typical district does not test before third grade. Thus, we can only use data only on students who switch after the fourth grade or later. This is highly problematic for several reasons. First, as noted above, the self-selection problem is more intense among late grade

---

[13] We could also use data from the *vice versa* situation: switching from a charter school to a regular public school. To avoid clumsiness, *vice versa* may hereafter be assumed.

switchers than among students who enter their chosen school at kindergarten or first grade. This is both because the inertia effect grows and because the parents' private information grows dramatically relative to the information the researcher observes. Intuitively, if there are two children who make identical gains between the third and fourth grades but one child is pulled out of the school where he knows his peers, the building, the culture, and the curriculum, then the two children are *necessarily* very different on unobservables. A researcher is not only not curing selection bias, he is discarding children for whom selection bias is not severe and is focusing exclusively on children for whom selection bias is severe. Since the bias is based on unobservable variables, he can do nothing about it even though he knows he has exacerbated it.

One might think that this problem can be mitigated greatly by earlier testing. Perhaps if all kindergarteners and first graders were tested, value-added analysis could work with children who switch as early as second grade? Unfortunately, this is not the case because kindergarten and first grade tests have much less predictive power than third and fourth grade tests. The assertion on which value-added is founded is that we can forecast future achievement with past achievement. We do not merely need test scores, therefore; we need test scores that are highly informative about future achievement.

The second problem with value-added analysis is power. The researcher is throwing away most of the data--typically, all of the observations on students who enter charter schools before fourth grade. He will thus find it hard to identify effects of policy-relevant size simply because his sample is so small and his standard errors are so high. He is likely to produce estimates that are so noisy that substantial positive, substantial negative, and zero effects all remain likely. That is, positive effects of policy-relevant size will be in the confidence interval but so will be zero and negative effects. Such estimates are virtually useless. Although people

sometimes misinterpret noisy estimates and say that they mean that the effect is zero, such an interpretation is strictly wrong and should be firmly squelched.

The power problem is serious because it is not tractable. One might think that one can escape it simply by waiting for more years of data on the charter schools in question. Waiting does not help, however, because, as charter schools become more established, they admit fewer and fewer students in grades that are not typical entry grades. That is, in its start-up year, a charter school may admit students in every grade that it offers--kindergarten through sixth grade, say. In the first year of admission, there will be a certain number of students (those who enter in the fifth grade) on whom value-added analysis can be performed. Unfortunately, that number will not be much amplified by successive years of operation, in which the charter school admits a student to its upper-level grades only if some student has vacated a seat. There will be just a trickle of late grade entrants after the first couple of years, and the power problem will not disappear.[14] Moreover, even if one can get sufficient power by relying on data just from charter schools' start-up years, estimates generated in this way are highly problematic. Start-up years are not representative of charter schools in general so that the estimates based exclusively on them are hard to interpret and impossible to extrapolate. In addition, as discussed above, self-selection problems are at their most intense in start-up years when parents who choose a charter school must accept a high level of uncertainty.

The final problem with value-added analysis is that a child's prior achievement is endowed with tremendous importance because it alone is used to predict future achievement.

---

[14] Indeed, because the new students enter in different years than the remainder of their classmates, they absorb degrees of freedom (for years of treatment by grade effects, for instance) even as they add observations. Their net contribution to power is therefore negligible.

Yet, we do no have the child's true prior rate of gain, we merely have an erroneous measure of it. This error will, at a minimum, exacerbate power problems; it may cause bias as well.[15] In order to substantially reduce the measurement error in an equation like (1), we would need four observations of prior test scores. (Four observations allows us to average two independent measures of gain.[16]) However, a researcher who requires four prior test scores can only use data on students who switch after the seventh grade! In other words, we cannot do anything about the measurement error without greatly exacerbating the selection bias that arises from relying on late switchers.

To demonstrate some of the facts we mention above, we offer Table 1, which shows the number of students who apply to charter schools in each grade of entry in Chicago and New York City.

Table 1 here

Table 1 shows that the vast majority of students enter charter schools in kindergarten or grade one: 55 percent of Chicago students admitted to charter schools enter in these grades and 46 percent of New York students do the same. Yet these data, if anything, exaggerate the extent of late-grade entry because both cities' charter schools include numerous recent start-ups. In New York, there is a slight bump up in entries in grade five (KIPP schools begin in that grade) and grade nine (when high schools begin), but otherwise the applications drop off with each grade. The data on grade of entry demonstrates just how odd are the switchers on whom value-

---

[15] Bias caused by measurement error in a multivariate setting is hard to predict, even supposing that parents do not react to the error in test scores but only react to their child's true achievement.

[16] There is only a trivial gain in forecasting power from adding only one observation (to make a total of three). This is because the middle observation is used twice and serial dependence, such as reversion-to-the-mean, makes the second gain highly dependent on the first.

added analysis exclusively relies. Value-added analysis forces a researcher to focus on a small sample, disproportionately plagued by selection bias.

These problems cannot be fixed simply by shifting to another value-added equation. A researcher who thinks that a student's prior *level* of achievement is sufficient for forecasting future achievement could include data on students who switch one year earlier: rising fourth graders might be added to the estimation sample. However, specification tests typically reject the hypothesis that the achievement level is sufficient for predicting future achievement, and reversion-to-the-mean becomes a serious problem that is impossible to address with only one year of prior data. A researcher who decides that a good forecast requires an equation of form (2) may be correct as to modeling, but must rely on even later switchers than a researcher who uses an equation of form (1). He is exacerbating the problems of selection bias and power associated with his ever-dwindling estimation sample.

In summary, value-added analysis is not a useful method for generating evidence on charter school effects. It does not credibly reduce the problems of selection bias that also plague comparison-with-controls methods. Instead, it exacerbates selection biases and reduces the power of the estimates. Unlike comparison-with-controls methods, where better judgement (however scarce) may improve the estimates, better judgement is not helpful with value-added analysis. The problems specific to value-added analysis are intrinsic: they stem from the essential nature of the decision to send a child to a charter school. The unobserved force that causes otherwise identical parents to apply to a brand-new charter school (susceptible to value-added analysis) is larger than the force needed to make them apply to a mature charter school (not susceptible to value-added analysis). As a child progresses in his school career (and becomes susceptible to value-added analysis), the unobserved event that will shock his other

identical family into making a switch grows larger. Similarly, as a child progresses in his school career (and becomes susceptible to value-added analysis), the parents' informational advantage over researchers grows, thus reducing the likelihood that prior achievement is sufficient for predicting future achievement.

D. Lottery-Based Analysis

The foundation of lottery-based analysis is fundamentally different from that of either comparison-with-controls or value-added analysis. It fact, its foundation is purely statistical: randomizing students between two groups will produce groups that are balanced on both observable and unobservable characteristics if the sample being randomized is sufficiently large. That is, randomization over a large number of students who apply to a charter school eliminates all forms of self-selection bias, whether from observable variables such as prior achievement or unobservable variables such as parental motivation. This is a statistical property, not an assumption. Therefore, an analysis that is based on comparing students who are lotteried-in and lotteried-out of a charter school requires only two things to generate useful, valid estimates: the lottery must be actually be random and the number of students must be large enough to generate balanced samples. Fortunately, we can do more than hope that these requirements are met: they can be checked. (We will return to this point.) If the two requirements are met, all other problems associated with self-selection are tractable. This is not to say that there are no data or other concerns that affect lottery-based analysis, but the remaining concerns equally affect comparison-with-controls and value-added analysis and have not been of sufficient importance to enter our discussion thusfar.

Another way of understanding why lottery-based analysis solves selection problems of all types is that all children in a lottery have already selected to apply to a charter school.

Comparison-with-controls and value-added methods always make us ask, "If these children and their families are so alike, then why are they making different choices about an important thing like the school they attend?" The lottery-based method eliminates all such questions. The children and their families *are* making the same choices; a random number is turning the same choices into different school experiences.

Because lottery-based analysis computes effects using the difference between lotteried-in and lotteried-out applicants, its evidence has a straightforward interpretation. The effect computed is the effect we should expect the charter school to have on students who wish to attend and who therefore go through the application process. Since charter schools are strictly voluntary, this is exactly the estimate we need for policy purposes.

Ideally, lottery-based analysis proceeds as follows. Students apply to a charter school, and it holds a random lottery among the applicants. The lottery should naturally be held in a fashion that generates *prima facie* validity (for instance, drawing chits from a box or using a random number generator), but the randomization can also be checked *ex post* by testing whether the lotteried-in and lotteried-out students who participated in the same lottery have observable characteristics that differ statistically significantly. This is not a full check, of course, but it can be made arbitrarily extensive using "no-stakes characteristics" such as the number of times the letter "k" occurs in a child's street address.[17] If a lottery is random but too small relative to the

---

[17] No-stakes characteristics are variables that are readily available for students in the lottery but that are not supposed to have the power to predict achievement. For instance, the number of times that the number 3 appears in a student's home address is a no-stakes characteristic. As the number of characteristics being checked for differences becomes large, we should find that the percentage of characteristics that appear to differ statistically significantly converges on the level of statistical significance being employed. We can make the number of characteristics being checked arbitrarily large by using no-stakes characteristics in addition to those that are expected to have explanatory power: race, gender, and so on.

variation in the population, it may generate random but unbalanced lotteried-in and lotteried-out groups. Thus, the next check is for balance on the observable characteristics of students. Again, this is not a full check because we cannot actually check unobserved variables. Nevertheless, the check can be made arbitrarily extensive using no-stakes characteristics. Furthermore, we describe another, and more important, check for balance below.

In our Chicago and New York City charter school evaluations, we collect each charter school's lottery information--that is, the set of applications along with the lottery number assigned to each student. Nearly always, lotteries are grade-specific for each school, a point that is important to remember. Schools do not always hold lotteries for every grade. A start-up school might, for instance, have lotteries for its kindergarten, first grade, second grade, and third grade, but be able to accommodate all applicants for its fourth, fifth, and sixth grades. We collect application information even if a lottery was not held. Later, we discuss how we can use such information. Our Chicago sample includes nine charter schools which together enroll most of the charter school students in the district. Our New York City includes 42 of the 46 charter schools in operation by 2005-2006. The application information is matched to the district's database, and information is retrieved not only on a student's post-lottery achievement (used to compare outcomes) but also on his pre-lottery characteristics, including his prior achievement scores if any. Pre-lottery data not only allows us to check for randomness and balance, it also allows us to investigate whether a charter school draws applicants who are particularly high or low achievers relative to other local students. In Hoxby and Rockoff (2004) and forthcoming reports on New York City charter schools, we show checks for randomness and balance. We have never found evidence of a non-random lottery, but we often find a lack of balance in small lotteries held for seats in grades that are not typical entry grades. For instance, if there are two

sixth grade seats available and twenty students participate in a lottery for the two seats, the lotteried-in and lotteried-out groups typically are not balanced. Later, we discuss how information from unbalanced lotteries can be used, but unbalanced lotteries cannot be used generate lottery-based estimates.

Lottery-based estimates are generated by computing the lotteried-in versus lotteried-out difference in achievement for each post-lottery year, for each lottery that did not fail the balance criteria. This leaves us with a large number of estimates--one for each group of students who participated in a lottery for a particular grade in a particular school in a particular year. Such a large number of estimates is not terribly informative, so we compute aggregates of them that we believe to be interesting. For instance, if one wants to know just the average effect of charter school for a student in the sample, one takes a weighted average of all the lottery-specific effects, where the weights are the number of students used to compute each effect. If one wants to know the average effect of charter school for a student who enters in the first grade, one takes a weighted average of the lottery-specific effects in which the first grade is the entry year. If one wants to know the effect of attending a charter school that has been in operation for three years when the student enters, one takes a weighted average of the lottery-specific effects in which the entry is the school's third grade of operation. And so on. Because one cannot show all aggregates, it makes sense to show those that are most relevant to policy. Experienced researchers will realize that we do not actually compute effects for each lottery and then aggregate them later. Instead, we compute the desired aggregate(s) while simultaneously computing individual lottery fixed effects. It is easier to calculate the correct standard errors for the aggregates using simultaneous estimation.

The only aggregates we cannot compute reliably are those for which we lack an adequate

number of balanced lotteries. This brings us to the second, and more important, test of balance. Randomization over a sufficient number of students automatically generates observed and unobserved characteristics that are balanced in a statistical sense. Therefore, the charter school effect we compute should not depend on whether we control for student characteristics or not. This is the beauty of randomization. It is also good intuition for why randomization obviates concerns about self-selection associated with unobservable variables. We can test whether a lottery is balanced by adding and subtracting observable characteristics from the estimation and seeing whether the estimated charter school effect varies to a statistically significant extent. In practice, we should not only carry out this test statistically, we should also see whether the estimated charter school effect varies to an extent that would confound a policy maker trying to make laws. If the charter school effect varies as we control for different sets of observable characteristics, then the lotteries on which it is based are insufficiently balanced to produce useful estimates.

Not all students who are lotteried-in (offered a place) when they participate in a charter school lottery actually take the place they are offered. This occurs most often, in practice, because a family has moved or its circumstances have changed, because a sibling who also applied was lotteried-out, or because the student was simultaneously offered a place in another charter or magnet school. Regardless of the reason, a student who is lotteried-in was "intended to enroll" at a charter school. When social scientists and medical doctors perform lottery-based analysis, it is standard to compute not just the effect of actual enrollment, but also the effect of being intended for enrollment. In our studies, we compute "intention-to-treat" results for completeness, but the effect of being intended for enrollment has little or no policy relevance in the charter school setting. This is because attending a charter school is always meant to be

voluntary. Put another way, a student who does not comply with the lottery's "intention" for him is not doing anything wrong; he is not like a patient who does not comply a doctor's intended treatment. The effect that *is* relevant for policy is the selection-bias-free effect of attending a charter school (the "treatment on the treated effect"). Because some students who are lotteried-in do not actually attend, we compute this effect by instrumenting for the indicator for attending charter school ($I_{it}^{charter}$) with an indicator for the lottery "intending" the student to attend charter school ($I_{it}^{lotteried-in}$). In fact, all along we have been describing these treatment-on-the-treated effects as the charter school effects computed based on the lotteries. We have delayed describing the instrumental variables procedure purely for expositional purposes.

An interpretative issue for lottery-based analysis is late offers. A late offers occurs when a place opens up a charter school after the lottery (usually held in the spring). The charter school then takes out its list of applicants and contacts the student whose lottery number was just below the threshold for being lotteried-in. If he does not accept the place, it is offered to the student with the subsequent lottery number, and so on. The reason that late offers pose an interpretation problem is that a parent who has meanwhile made other arrangements for his child may refuse a late offer even if he would have accepted an on-time offer. Thus, it is not clear that we ought to lump on-time offers and late offers together in a single indicator and assume that they have the same effect. Rather than attempt a definitive answer to this interpretation problem, we prefer to show estimates of charter school effects that are for on-time offers only and for all offers.[18] In

---

[18] Formally, there are three possible indicators for being lotteried-in: being lotteried-in at the time of the lottery ($I_{it}^{lotteried-in\ on-time}$), being lotteried-in late ($I_{it}^{lotteried-in\ late}$), and being lotteried-in regardless of timing ($I_{it}^{lotteried-in}$, which is the sum of the previous two indicators). We compute estimates using the first and third indicators as instrumental variables. There are an insufficient number of students who are lotteried-in late to compute a separate estimate for them.

practice, these effects have turned out to be so similar that it does not seem worthwhile wrestling further with the interpretation problem.

When it is available, lottery-based analysis is strictly superior to comparison-with-controls methods or value-added analysis. Although some estimation problems, such missing data and attrition (students who are lost to data cachement altogether because they, say, move out of the district), pose problems for all three methods, lottery-based analysis solves all selection problems much more credibly than the other methods and does not aggravate any estimation problems. Thus, when they are available, lottery-based estimates are the "gold standard".

E. Combining Methods for Maximally Comprehensive Evaluation of Charter Schools' Effect on Achievement

The creation of a gold standard has, for some time, suggested a remedy for the one deficiency of lottery-based analysis: the fact that it is not always available. The small lotteries for atypical entry grades are often not balanced and therefore not susceptible to lottery-based analysis. Some charter schools can accommodate all applicants in their first year or two of operation and only hold lotteries after they have been in business for a few years. Some charter schools, particularly those located in rural areas, almost never hold lotteries.

The best way to illustrate the use of the gold standard is to provide a practical example. In our work, we often find unbalanced lotteries for students who apply to atypical entry grades. Such students are susceptible to comparison-with-controls analysis. The difficulty with comparison-with-controls is that it requires considerable judgement from the researcher, and even a researcher who exercises brilliant judgement may have that judgement questioned by others. The availability of gold standard estimates for the other grades-of-entry in a charter

school allows a researcher to test empirically his comparison-with-controls specification. That is, he should be able to demonstrate that his specification closely replicates the lottery-based estimates for the data for which both are available. Whether a sufficiently accurate comparison-with-controls specification can be found is, of course, an empirical matter.

Readers may derive intuition from an alternative logic about the relationship between comparison-with-controls and lottery-based analysis. Lottery-based analysis works because there is a group of students outside the charter school (the lotteried-out) who are good controls for the students attending the charter school. Lotteries make it easy to find the appropriate control group of students but, in return for this ease, they do not always identify appropriate controls even when they exist. This is because lotteries depend on the law of large numbers and have no special way of picking the appropriate controls out of the crowd when the lottery is not large enough to provide balance. In contrast, comparison-with-controls methods create a rule for picking appropriate controls out of the crowd. We can even think of a superb comparison-with-controls specification picking out exactly the lotteried-out students without knowing that they had participated in the lottery. Therefore, a comparison-with-controls specification can, in principle, pick out the appropriate control for a charter school student who, say, participated in a lottery too small for balance.

In the event that a researcher is able to identify a comparison-with-controls specification that replicates his lottery-based results, he may then reasonably speculate that if he applies the same comparison-with-controls method to the data on *all* students who attend the charter school in question, he will obtain fairly credible, comprehensive results.

So far, our example has been grades with small lotteries, but using lottery-based results to validate a comparison-with-controls specification may also work for a charter school that does

not hold a lottery in its first year of operation but thereafter does. Far more caution is needed if a researcher wishes to use lottery-based results from one school to validate a comparison-with-controls specification that he plans to apply to other schools. Extrapolating from one school to another is a speculative exercise unless there are *a priori* reasons--such as one school being a branch or offshoot of the other--to believe that the schools are inherently similar.

In short, comparison-with-controls methods (including those that use gain scores) are deficient because it is difficult to verify whether they have solved selection problems when they are used. Indeed, such methods may be unable to solve selection problems even if used well, and they may aggravate selection problems if used poorly. The availability of lottery-based estimates may allow a researcher to address this deficiency empirically, thereby opening the door for estimates of charter school effects that are comprehensive--that is, estimates based on every charter school student's achievement. Combining lottery-based and comparison-with-controls methods is a sound overall strategy for generating comprehensive estimates, so long as researchers are aware that comprehensive estimates may remain unobtainable. We are better off having only credible lottery-based estimates than having a mixed bag of credible lottery-based estimates and non-credible comparison-with-controls estimates.

We can think of no justification for mixing value-added analysis with the other methods for analyzing charter school effects. This is because the problems with value-added analysis are inherent. Put another way, demonstrating that value-added analysis and lottery-based analysis produce similar estimates for students for whom they are both available does not allow a researcher to argue by extension that value-added analysis is valid for those students for whom lottery-based analysis is unavailable. It may be easiest to explain this point with a concrete example. KIPP schools usually cover grades five through eight. Therefore, they admit most of

their students in grade five lotteries. A typical KIPP school also enrolls a trickle of students to its later grades. Suppose that a KIPP school has a large fifth grade lottery and operates in a district where testing occurs in all of grades three through eight. Then, its fifth grade entrants will be susceptible to both lottery-based and value-added analysis. Even if the two methods produce similar results for the fifth grade entrants, one cannot conclude that value-added estimates for later grades will not suffer from substantial selection bias. This is because value-added analysis exacerbates selection bias precisely for students who enter in atypical grades. The fact that the value-added analysis on the typical grade entrants (grade five for KIPP schools) did not produce much selection bias tells us nothing about whether value-added analysis will suffer greatly from selection bias in atypical grades.[19]

IV. Some Illustrative Estimates from Our Evaluations of Chicago and New York City Charter Schools

This is primarily a methodological paper, but readers may find it helpful to consider a few concrete results that illustrate some of the points discussed above. For more extensive results, see Hoxby and Rockoff (2004) and forthcoming reports on New York City charter schools. [Chicago results are shown here, but New York City results are unfortunately embargoed until the end of October 2006. New York City results will be added to this paper as

---

[19] Readers may be interested in a subtle distinction here between comparison-with-controls and value-added. Value-added analysis always compares a student to his former self. This comparison is problematic for atypical grade entrants precisely because some event must have precipitated them to enter in an atypical year. We cannot credibly argue that the same event has no effect on later achievement. In contract, comparison-with-controls methods seek the right control students among the possible controls. In principle, there could be some control student who experienced an equivalent event and who can therefore generate unbiased estimates. Finding the appropriate control student may be difficult, but it is at least possible.

soon as they may be published.]

Table 2 shows computations of a simple average charter school effect, separately for Chicago (left panel) and New York City (right panel). To compute the results, we included all balanced lotteries. The basic result is shown in the left-hand column of each panel: the average treatment-on-the-treated effect of attending a charter school. The average effect of charter school attendance is about 6 percentile points in math and about 5 percentile points in reading for Chicago students. The average Chicago student in the data has attended charter schools for two years, so one-year effects in math and reading are 3 percentile points and 2 percentile points, respectively. The average effect of charter school attendance is [] in reading and [] in math for New York students.. The average New York City student in the data has attended charter schools for [] years, so a one-year effect in reading and math are [] and [], respectively. The fact that the lotteries are balanced is demonstrated in the second and third columns of each panel. There, we control for various sets of observable student characteristics, and the estimates of the charter school effect do not change to a statistically significant effect. Moreover, they do not change to an extent that would be relevant for policy makers. In the fourth column of each panel, we show results based only on all students who are lotteried-in, not just those who receive on-time lottery offers. This change creates differences in the results that are neither statistically significant nor of a size relevant for policy making. The similarity of the estimates for on-time and all lottery offers suggest that families who receive late offers react in the same way that they would have reacted had they received an on-time offer.

Table 2 here

Table 3 deliberately shows invalid results based on lotteries that are too small to be balanced. Otherwise, Table 3 has a similar format to Table 2. Notice that the number of

observations in Table 3 is much smaller than that in Table 2. This is because the vast majority of applicants to the charter schools are in balanced lotteries. Any large flow of students into a typical grade of entry will normally end up in a balanced lottery. Therefore, as a fairly mechanical matter, most students will be in balanced lotteries. Notice also that when we control for various sets of observable student characteristics, in the second and third columns of each panel, the estimated charter school effects jump around to an extent that would be very problematic for policy makers. Put another way, when selection bias is present, as it is when lotteries are too small to balance, estimated effects are not reliable. This lack of reliability is not something that standard errors can reveal because it is result of uncontrolled biases, not merely noise in the data.

Table 3 here

Finally, Table 4 shows a comparison of lottery-based and value-added results for Chicago (left panel) and New York City (right panel). In each panel, the left-hand and middle columns both show lottery-based results, but the left-hand column shows the lottery-based result that uses all of the balanced lotteries. The middle column, in contrast, uses only the *much smaller* sample of students whose achievement is susceptible to value-added analysis. This sample of late grade movers is less than one-tenth the size! The late grade movers also appear to be an unusual bunch: the lottery-based estimate for them has a large standard error and is quite different from the estimate based on all students (although not statistically significantly different owing to the large standard error). In the right-hand column of each panel, we show the value-added estimate for the small sample of late grade movers who are susceptible to value-added analysis. For each city, this estimate is radically different than the gold-standard lottery-based estimate, indicating the amount of selection bias present in--even exacerbated by--value-added

analysis. We conclude that value-added analysis is not a reliable method for estimating charter schools' effects on achievement.

Table 4 here

[We hope to add some estimates comparing lottery-based results with the best comparison-with-controls specifications that we are able to obtain. However, it remains to be seen whether we will be able to find specifications that test well empirically.]

V. Learning What Works from Charter Schools

As we noted at the outset, a policy maker may reasonably want evidence on what works in public education and may view charter schools as laboratories in which interesting educational experiments occur. In order to deliver this kind of evidence, a researcher must first compute credible estimates of each charter school's effects and then relate them to charter school characteristics. As we have seen, computing a credible estimate of each charter school's effect is a significant challenge and may not be possible in the case of new schools, small schools, and schools that do not run substantial lotteries. Assuming, however, that such estimates have been created, the next logical step is using multiple regression or a similar method to relate effects to school characteristics that are quantifiable or at least classifiable. There can be little doubt that researchers aspire to be able to provide this sort of evidence, and there is a great deal of pressure on them to produce it. Unfortunately, producing such evidence is not as easy as many policy makers believe. In this final section of the paper, we discuss the promise and pitfalls of this sort of research.

Consider Table 5, which shows seventeen quantifiable or classifiable characteristics of New York City charter schools in our sample. The first thing to observe is that there are some

substantial educational experiments taking place in the schools. 89 percent have a long school day, with a few having days of eight hours or more. 62 percent have a long school year, with some having years with more than 200 days. In about half of the schools, children attend Saturday school and study language arts (reading) for more than one hour per day. About three-quarters require children to wear uniforms, 65 percent regularly administer internal assessments (in addition to the state tests), and 30 percent offer substantial merit pay to teachers. Table 5 also shows some management types and curricula used. In addition to the characteristics shown, we could easily observable variables such as the number of years a school has been in operation; which New York authorizing organization granted the charter; whether the school has a dedicated school building, shares a building with a regular public school, or uses facilities not originally for a school.

Despite its length, the list in Table 5 contains only a small subset of the interventions about which researchers are asked to provide evidence. Yet, we have selected the list with some care to include the educational strategies that are used by multiple schools and that can be quantified or classified. (We allow the school themselves to classify their strategies if the school description is unclear. Thus, it is only strategies that the schools themselves find impossible to classify that are excluded on this basis.) Policy makers who ask for evidence are sometimes unaware of the fact that we *cannot* separately identify the effect of a strategy employed by a single school and we are very unlikely to be able to separately identify the effect of strategy employed by only a few schools. The reason we must exclude strategies that resist classification is they are so dependent on implementation that they cannot be separately identified from particular management teams.

Another noteworthy thing about the list in Table 5 is that we reduced the detail on

various characteristics. For instance, rather than include a identifier for each Charter Management Organization (CMO), we have lumped them together. Rather than include information on the detail of each merit pay plan for teachers, we have lumped them together. This is because it is necessary to reduce detail to ensure that a characteristic is not associated with a single or only a couple of schools (in which case its effect cannot be separately identified). Moreover, even though we do not have the space to show a full correlation matrix, it is the case (and should be fairly evident) that collinearity among certain characteristics is likely to be a issue. For instance, we do not find the long school year or the long reading period in schools that do not also employ a long school day. The prevalence of CMOs and Educational Management Organizations (EMOs) ensures that some characteristics come in "clumps".

Summing up: individual charter school effects are estimated with error; school characteristics tend to be collinear; even a very modest list of characteristics (with deliberately reduced detail) numbers one third the number of charter schools in New York City and approximately one-half the number for which an individual school effect can be estimated at this time. It is evident that we face a formidable estimation challenge. Moreover, we have so far not mentioned the fact that, even in New York City which has a large number of charter schools in a compact area, there is substantial variation in charter schools' locations, buildings, and student demographics. These variables could easily have independent effects or effects that interact with the strategy variables listed above. For instance, researchers are often asked whether school uniforms work equally well with black students as with Hispanic students (school uniforms are traditional in many of the home countries from which Hispanic students come). Researchers are asked whether particular disciplinary policies work better when a school has a stand-alone building than when it shares a building with a regular public school (which may not have the

same disciplinary policy). In short, the fact that charter schools operate in different environments means that researchers must be cautious about ascribing achievement effects to school strategies unless they have, at a minimum, simultaneously controlled for a number of school environment variables. The estimation problem only gets more difficult when we acknowledge that strategies are endogeneous to a school's environment–for instance, a long school year might be much more popular with single parent families than with others or a parent contract may be infeasible in a school that parents do not get to choose. Differences in environment also limit the benefits associated with increasing the sample to, say, include charter schools in New York state but outside New York City. A charter school in Utica (a small in upstate New York) might add variation in school effects and school characteristics, but it would also add such a large amount of variation in school environment that its contribution to evidence on "what works" would be small at best.

We conclude with the hope that the effects of characteristics like those shown in Table 5 may be identifiable, but we also conclude with caution regarding the immediate promise of charter school research for discovering "what works" in education. Until we have dramatically greater availability of gold standard evidence on individual charter schools' effects, claims about identifying the effects of educational interventions should be extremely modest.

Table 1
Student Admissions by Grade

	Chicago Charter Schools	New York City Charter Schools
	Grade accounts for this percentage of all admitted students	Grade accounts for this percentage of all admitted students
KG	45.4%	30.8%
Grade 1	9.8%	15.4%
Grade 2	7.1%	10.1%
Grade 3	6.3%	7.4%
Grade 4	7.0%	5.9%
Grade 5	6.9%	13.8%
Grade 6	6.5%	4.7%
Grade 7	6.3%	1.7%
Grade 8	4.7%	0.8%
Grade 9	~0%	4.6%
Grade 10	~0%	4.1%
Grade 11	~0%	0.6%
Grade 12	~0%	0.1%

Table 2
The Average Effect of Attending Charter School on Math and Reading Achievement
Treatment on the Treated Effects

	Chicago				New York City			
	Basic estimate	Control for observable pre-lottery student characteristics (except prior achievement)	Control for observable pre-lottery student characteristics including prior achievement	Use all lottery offers (not only on-time offers)	Basic estimate	Control for observable pre-lottery student characteristics (except prior achievement)	Control for observable pre-lottery student characteristics including prior achievement	Use all lottery offers (not only on-time offers)
Effect on Math Percentile Score	5.61 (2.85)	5.66 (2.65)	6.20 (2.10)	6.31 (2.45)				
Effect on Reading Percentile Score	4.82 (2.78)	5.00 (2.57)	5.33 (2.00)	5.97 (2.36)				
Observa-tions	2701	2701	2701	3060				

Notes: See Hoxby and Rockoff and forthcoming New York City reports for details regarding schools and data. All estimations include a full set of estimated effects for individual lotteries (which are grade and school specific). The aggregate effects shown also include grade-by-year effects to account for variation in the tests from year to year and from grade to grade. The average student in the Chicago sample had attended charter school for 2 years. The average student in the New York City sample had attended charter school for [] years.

Table 3
Invalid Estimates Based on Unbalanced Lotteries
Invalid Treatment on the Treated Effects of Charter Schools on Math and Reading Achievement

	Chicago				New York City			
	Basic estimate	Control for observable pre-lottery student characteristics (except prior achievement)	Control for observable pre-lottery student characteristics including prior achievement	Use all lottery offers (not only on-time offers)	Basic estimate	Control for observable pre-lottery student characteristics (except prior achievement)	Control for observable pre-lottery student characteristics including prior achievement	Use all lottery offers (not only on-time offers)
Effect on Math Percentile Score	-4.30 (3.68)	-7.24 (3.63)	-4.78 (1.83)	-2.47 (4.10)				
Effect on Reading Percentile Score	-0.87 (3.10)	3.44 (3.06)	-3.11 (1.81)	-1.18 (3.57)				
Observa-tions	573	573	573	649				

Notes: See Hoxby and Rockoff and forthcoming New York City reports for details regarding schools and data. All estimations include a full set of estimated effects for individual lotteries (which are grade and school specific). The aggregate effects shown also include grade-by-year effects to account for variation in the tests from year to year and from grade to grade. The average student in the Chicago sample had attended charter school for 2 years. The average student in the New York City sample had attended charter school for [] years.

Table 4
Lottery-Based and Value-Added Estimates, Compared
Effects of Charter Schools on Math and Reading Achievement

	Chicago			New York City		
	Lottery-based estimate for all students in balanced lotteries	Lottery-based estimate that relies solely on students susceptible to value-added analysis	Value-added estimate that relies solely on students susceptible to value-added analysis	Lottery-based estimate for all students in balanced lotteries	Lottery-based estimate that relies solely on students susceptible to value-added analysis	Value-added estimate that relies solely on students susceptible to value-added analysis
Effect on Math Percentile Score	6.31	-1.38	1.03			
	(2.45)	(5.92)	(7.02)			
Effect on Reading Percentile Score	5.97	-0.75	-1.35			
	(2.36)	(5.29)	(8.64)			
Observations	2701	562	562			

Notes: The two methods are described in the text. The sample in the two right hand columns of each city's panel is limited to students for whom value-added estimates can be computed. All students who apply to charter schools in balanced lotteries are included in the estimates shown in each panel's left-hand column. The lottery-based estimates are treatment-on-the-treated effects.

Table 5
Some Characteristics of New York City Charter Schools

School characteristic or practice	Share of New York City charter schools
Operated by a Charter Management Organization (CMO)	27%
Operated by an Education Management Organization (EMO)	17%
Operated by a Community Grown Organization (CGO)	56%
Long school day	89%
Long school year	62%
After-school program available	56%
Saturday school (mandatory for all or certain students)	51%
Long English/language arts period (one hour or more)	54%
Saxon Math curriculum	19%
Open Court Reading curriculum	11%
Core Knowledge curriculum	22%
Internal assessments regularly administered	65%
Parent contract	38%
"No broken windows" discipline philosophy	22%
Uniforms required	76%
Teachers unionized	14%
Merit pay or bonuses for teachers	30%

Notes: This table is based on thirty-seven New York City charter schools that have provided and confirmed their descriptive information.